Mapping of Proteinase Active Sites by Projection of Surface-Derived Correlation Vectors

MARTIN STAHL, DANIEL BUR, GISBERT SCHNEIDER

F. Hoffmann–La Roche Ltd., Pharmaceuticals Division, Molecular Design and Bioinformatics, CH-4070 Basel, Switzerland

Received 21 August 1998; accepted 4 October 1998

ABSTRACT: A method for the mapping of surface cavities in proteins is presented. Cavities were defined by a grid-based accessibility measure, where a protein atom was considered to form part of the cavity surface if its accessibility value was below a critical threshold. Surfaces were calculated by the Connolly algorithm. To each surface point, one of five generalized atom types and an accessibility value were assigned. Based on this surface description, topological correlation vectors were generated and projected onto the plane by means of linear and nonlinear mapping techniques. The method was successfully applied to represent relationships among a diverse set of aspartic, serine, and metalloproteinases. © 1999 John Wiley & Sons, Inc. J Comput Chem 20: 336–347, 1999

Keywords: active site; protein cavity; topological correlation; nonlinear projection; neural network

Introduction

The increasing number of experimentally solved three-dimensional (3D) structures of proteins enhances the need for automated techniques to analyze, classify, and compare these structures. In recent years, many computational tools have been developed that aim at the identification of common subunits and 3D motifs in proteins.[1] Such tools can provide valuable information for a better understanding of protein architecture and structural building blocks. From a pharmaceutical point of view, crucial information is encoded in the surface properties of a protein, because inhibitors bind to specific pockets at the surface of an enzyme or receptor.[2] There clearly is a lack of fast and reliable methods for comparison and description of protein pockets, although there exist useful methods to detect surface cavities.[3]

Recently, a geometric hashing technique has been applied to match active site surfaces of various enzymes.[4] However, this method includes a

Correspondence to: M. Stahl; e-mail: martin.stahl@roche.com

CCC 0192-8651 / 99 / 030336-12

3D superposition of active sites, which is feasible for highly similar surfaces only. To overcome this limitation, it would be very convenient to have a technique that is independent of the spatial orientation of protein cavities, and that could be applied to closely as well as remotely related surfaces. For small druglike molecules, methods already exist that fulfill these standards.[5] Here, we describe such a mapping method for protein cavities and its application to the discrimination of proteinase active sites. Our method consists of three steps that can be considered independent of one another:

1. Generation of active site surfaces and assignment of surface properties.
2. Encoding of these surfaces by a vectorial descriptor representation.
3. Projection of the high-dimensional descriptor space onto a plane for visual analysis.

Each step introduces a new level of abstraction, but should nevertheless retain a maximum of information. In step 1, it is important to find general criteria for the definition of cavity boundaries. Furthermore, one must find descriptors that can be assigned to each surface point to describe essential cavity properties. We have approached both problems by using generalized atom types and an ''accessibility'' measure. Subsequently, topological correlation vectors were generated for each set of surface points describing an active site (step 2). Relationships between these vectors were analyzed and visualized by means of principal component analysis (PCA), nonlinear mapping (NLM), and self-organizing maps (SOM) (step 3).

To get insight into the strengths and weaknesses of our method, four sets of well-resolved x-ray structures of serine, aspartic, cysteine, and metalloproteinases were selected from the Brookhaven protein database (PDB)[6] for the purpose of evaluation (Table I). All selected enzymes are endopeptidases. The selection was done by one of us (D.B.) who was not involved in the development of the algorithm. Protein structures were selected after manual inspection on the basis of high diversity in shape and size. The catalytic residues that characterize the active sites of each type of proteinase are depicted in Figure 1. Because there is a high degree of variation among the members of each proteinase class, we believe that this test set is rather demanding for any clustering technique.

Computational Methods

ACCESSIBILITY MEASURE

A protein cavity is a region that is less accessible from the outside than other surface regions. This accessibility was quantified such that it could form the basis of the cavity definition algorithm. For an arbitrary point P it is calculated in the following way: A set of 45 points, distributed equally on a unit sphere,[7] is generated around P. Every line running through one of the sphere points and ending at P defines a direction along which access to P is either freely possible or hindered by the protein. The algorithm loops over these directions. Beginning at a value of 45, an integer variable is decremented if a ray passes through the van der Waals sphere of a protein atom. Only protein atoms within a distance of 15 Å from P are regarded. The resulting accessibility value, D, can assume values between 0 and 45. A value of $D = 0$ means that the grid point is completely buried within the protein, whereas a theoretical value of 45 would indicate that no protein is present. Values between 5 and 25 are typical for surface cavities of proteins. The large number of 45 directions is needed to ensure a maximum degree of isotropy of the accessibility measure; that is, only slight dependence on the spatial orientation of the protein.

CAVITY DEFINITION AND GENERATION OF ACTIVE SITE SURFACES

Starting from Cartesian coordinates of a point lying at the approximate center of the active site cavity, the algorithm finds all atoms that form the active site cavity. It then generates and processes a Connolly surface of these atoms.

The starting point of the cavity definition was the location of the catalytic zinc atom in the metalloproteinases, the side chain oxygen of the catalytic serine in serine proteinases, a carboxylate oxygen of one of the catalytic aspartates in aspartic proteinases, and the catalytic cysteine sulfur in cysteine proteinases (Fig. 1). All protein atoms within a distance of 17 Å from this point were included in the calculation. The following set of van der Waals radii was used: 1.7 Å for C; 1.6 Å for N; 1.5 Å for O; 1.8 Å for S; and 0.7 Å for Zn.

The algorithm consists of four formal computational steps that are schematically depicted in Fig-

TABLE I.
PDB Codes of Selection of Proteinases Used (Cysteine Proteinases Not Part of Training Set).

No.	Name		PDB Code	Res. (Å)
	Asp proteinases			
1	Renin	(*H. sapiens*)	2ren	2.5
2	Chymosin	(*Bos taurus*)	4cms	2.2
3	Pepsin	(*Sus scrofa*)	3pep	2.3
4	Cathepsin D	(*H. sapiens*)	1lyb	2.5
5	Endothiapepsin	(*E. parasitica*)	4er2	2.0
6	Proteinase A	(*S. cerevisiae*)	1jxr	2.4
7	SAP2	(*C. albicans*)	1eag	2.1
8	EIAV proteinase		1fmb	1.8
9	FIV proteinase		1fiv	2.0
10	HIV-1 proteinase		1hxb	2.3
11	HIV-2 proteinase		1hii	2.3
12	SIV proteinase		1siv	2.5
	Ser proteinases			
1	Thrombin	(*H. sapiens*)	1ppb	1.9
2	Trypsin	(*Bos taurus*)	1tnh	1.8
3	Chymase	(*Rattus norvegicus*)	3rp2	1.9
4	Subtilisin carlsberg	(*B. subtilis*)	2sec	1.8
5	Cathepsin G	(*H. sapiens*)	1cgh	1.8
6	Thermitase	(*T. vulgaris*)	1thm	1.37
7	Kallikrein	(*Mus musculus*)	1ao5	2.6
8	Proteinase A	(*S. griseus*)	2sga	1.5
9	CMV proteinase		1cmv	2.27
10	Toxin A	(*S. aureus*)	1agj	1.7
11	Tonin A	(*Rattus norvegicus*)	1ton	1.8
12	Chymotrypsin	(*Bos taurus*)	3gch	1.9
	Metal proteinases			
1	Thermolysin	(*B. thermoprot.*)	1tlp	2.3
2	Neut. collagenase	(*H. sapiens*)	1mmb	2.1
3	Stromelysin	(*H. sapiens*)	1hfs	1.7
4	Adamalysin	(*C. adamanteus*)	1iag	2.0
5	Matrilysin	(*H. sapiens*)	1mmq	1.9
6	Leishmanolysin	(*L. major*)	1lml	1.86
7	Serratia proteinase	(*Serratia sp.*)	1srp	2.0
8	Neutral proteinase	(*B. cereus*)	1npc	2.0
9	Metalloelastase	(*P. aeruginosa*)	1ezm	1.5
10	Atrolysin	(*Crotalus atrox*)	1dth	2.0
11	Astacin	(*Astacus astacus L.*)	1ast	1.8
12	Proteinase	(*S. caespitosus*)	1kuh	1.6
	Cys proteinases			
1	Gly endopeptidase	(*Carica papaya*)	1gec	2.1
2	Actinidin	(*Actinida chinensis*)	2act	1.7
3	Caricain	(*Carica papaya*)	1meg	2.0
4	ICE (CASP-1)	(*H. sapiens*)	1ice	2.6
5	Cathepsin B	(*Rattus norvegicus*)	1the	1.9
6	Cathepsin K	(*H. sapiens*)	1mem	1.8
7	Chymopapain	(*Carica papaya*)	1yal	1.7
8	Papain	(*Carica papaya*)	1pip	1.7
9	Proteinase omega	(*Carica papaya*)	1ppo	1.8

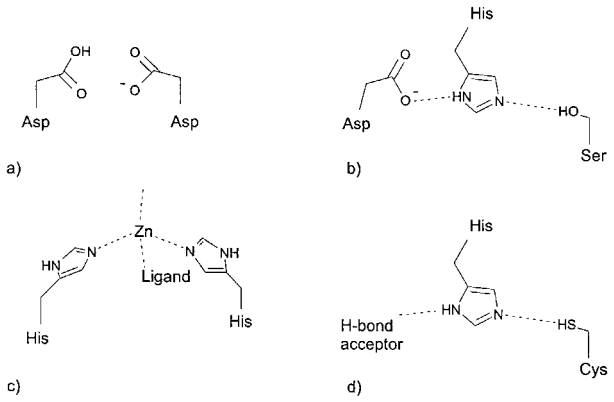

FIGURE 1. Schematic representation of the common catalytic residues in the four classes of proteinases examined herein: (a) aspartic; (b) serine; (c) metallo-; and (d) cysteine proteinases.

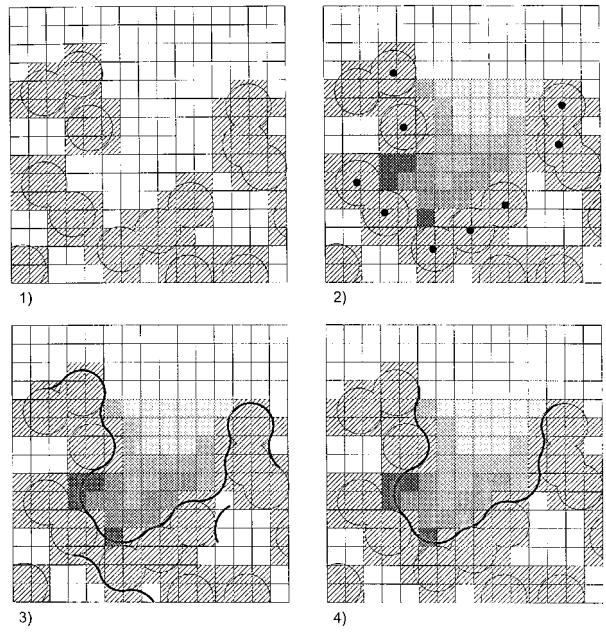

FIGURE 2. Two-dimensional scheme of the active site surface generation procedure. (1) A portion of the protein containing the active site is placed on a cubic grid of 1 Å spacing. In this representation, a grid point is located in the center of each square. A grid point is marked as occupied (hatched squares) if its center is less than 0.8 Å away from the van der Waals surface of a protein atom (open circles). (2) The active site cavity is defined by all unoccupied grid points having an accessibility variable, *D* (see text), that is below a certain threshold and connected with other points. A darker shade of gray symbolizes a lower value of *D*; that is, a less accessible region. The protein atoms forming the boundaries of the cavity are marked (filled circles). (3) The Connolly surface of these atoms is generated. (4) Double surface layers and disconnected patches are removed. See text for details.

ure 2. A cubic grid of 1-Å mesh size was placed on the protein atoms. A grid point was marked to be occupied by a protein atom if it was within a distance of 0.8 Å from the atom's van der Waals surface (Fig. 2, grid 1). For each of the unoccupied grid points, an accessibility value D was calculated. It was empirically found that a value of $D = 25$ is a sensible cutoff value for the outer limit of a protein cavity. Only grid points with values lower than 25 were defined to belong to a cavity. All connected grid points within the active site cavity were marked by a flood-fill algorithm (Fig. 2, grid 2). A Connolly surface[8] with a point density of six points/Å^2 and a solvent radius of 1.4 Å was generated of the protein atoms directly adjacent to this volume (Fig. 2, grid 3). Such surfaces often contain disconnected patches and regions bent away from the cavity. These artifacts yield a decreased accuracy in analysis, as was determined from test experiments. A clean-up routine removed all such surface patches, employing distance criteria between surface points and their corresponding closest cavity grid points (Fig. 2, grid 4). Finally, it was observed that some surface patches comprised fringes rather distant from the actual active site. To remove them, a spherical cutoff of 11 Å around the geometrical center of the surface was applied. Active sites of all enzymes included in this work (Table I) were contained within this volume. In a last step, each surface point was assigned an accessibility value D, and one of five generalized atom-type values T (hydrogen-bond donors, hydrogen-bond acceptors, aliphatic, aromatic, metal) according to the closest atom center. All sp2-hybridized carbon atoms in Phe, Tyr, Trp, and His residues were labeled aromatic. The remaining carbon atoms were considered to be aliphatic carbons. All nitrogen, oxygen, and sulfur atoms bearing hydrogen atoms were classified as donors, and all other polar atoms as acceptor atoms.

TOPOLOGICAL CORRELATION OF ATOM TYPES

Topological correlation vectors of atom properties can be used to obtain compact rotation- and translation-invariant descriptors of molecules.[9, 10]

This concept was applied to obtain descriptors of all active site surfaces. Correlation vectors, CV with components, CV_d^T, were generated as follows, based on the surface description of active sites in terms of accessibility values, D, and atom type values, T:

for each pairwise distance $d_{A,B}$ between grid points A and B{
  for each pair of atom types T_A, T_B, separated by distance $d_{A,B}${
    (in the case of autocorrelation) if $T_A = T_B$ then $CV_d^T = CV_d^T + D_A D_B$
    (in the case of cross-correlation) $CV_d^{T_A,T_B} = CV_d^{T_A,T_B} + D_A D_B$
  }
}

Distances between 0 and 10 Å were considered. This range was subdivided into n distance bins (Fig. 3), resulting in $5 * n$ dimensions for CV in the case of autocorrelation and $15 * n$ dimensions in the case of cross-correlation of atom types. Values of 5 and 10 for n were applied. This led to four sets of vectors (auto- and cross-correlation with $n = 5$ and $n = 10$), which were used for similarity analysis and clustering.

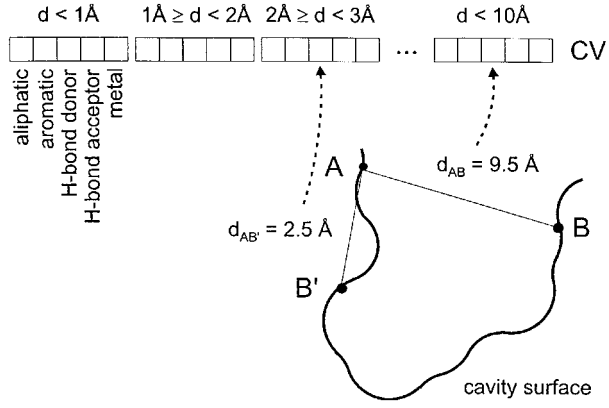

FIGURE 3. Generation of topological autocorrelation vectors of atom types. In this example, the surface points A, B, and B' are assumed to represent H-bond donors. The product of their accessibility values is added to the appropriate vector element.

PROJECTION DISPLAY OF MULTIVARIATE DESCRIPTORS OF PROTEASE ACTIVE SITES

Three projection methods were used to display the distribution of active sites in high-dimensional descriptor space. The algorithms applied produced linear (principal component analysis) and nonlinear (nonlinear mapping, self-organizing map) mappings of high-dimensional space. A thorough discussion of advantages and limitations of these techniques can be found elsewhere.[11,12]

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA rotates the vector (descriptor) space using the eigenvectors (the principal components) of the covariance matrix as a new basis.[13] The principal components are orthogonal, and their usefulness for data mapping can be assessed by the respective fraction of data variance explained (as expressed by their eigenvalues). The plane spanned by the two principal components with the largest eigenvalues was selected for data projection and display. PCA was performed using the program TSAR 3.1.[1]

[1]TSAR 3.1, Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Oxford, UK.

NONLINEAR MAPPING (NLM)

The principle of NLM[15] is to generate a two-dimensional display conserving relative distances between the N data points in high-dimensional space. NLM employs an optimization procedure in which the mapping error (E) is used as a quality criterion. E is a mean-square-error calculated as the difference between the interpoint distance matrices in high-dimensional space, $\mathbf{d}$, and the two-dimensional projection (Euclidian distance), $\mathbf{d}'$:

$$E = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (d_{i,j} - d'_{i,j})^2 \rightarrow \min$$

Optimization of the two-dimensional data coordinates was performed using a simple evolutionary

algorithm.[16,17] In a cyclic variation/selection-of-the-best scheme, the (x, y) coordinates in the 2D projection were optimized by minimizing E. (x, y) coordinates were randomly initialized in $[-1,1]$. In each optimization cycle (or ''generation''), 50 variant sets of 2D coordinates were generated, approximately gaussian distributed, around a set of ''parent'' coordinates, and the set leading to the lowest value of E was selected as ''parent'' for the next generation. The optimization was stopped after 200 generations.

SELF-ORGANIZING MAP (SOM)

Kohonen's SOM[18] algorithm belongs to the field of artificial neural networks.[19] One of its applications is visualization of chemical data.[12,20] A SOM is formed by a grid of formal neurons. We used a toroidal grid of neurons (an ''endless plane'') to avoid clustering artifacts arising from a finite planar topology. Due to the toriodal shape of the maps, each neuron has the same number of neighbors. Each neuron is characterized by its position (x, y) and a vector with the same dimension as the data vectors. The principle of SOM generation is to associate each data vector with a neuron. Data vectors that are close to each other in the high-dimensional space are mapped onto adjacent neurons. As in the NLM procedure this leads to nonlinear data projection, but, in contrast to NLM, map information about the actual interpoint distances in high-dimensional data space is lost. However, each neuron can be interpreted as a cluster grouping together data vectors that are most similar to each other, thereby introducing a classification scheme that is based on the topology of high-dimensional data space.

Algorithms for the generation of cavities, surfaces, and descriptors (except for the Connolly surface program), as well as the software for nonlinear projection, were developed in-house as a series of C modules.

Results and Discussion

A set of 36 diverse proteinase active sites (12 aspartic,[21] 12 serine,[22] 12 metallo[23]; Table I) was encoded by surface-derived correlation vectors. The distribution of the active site descriptors was visualized by mapping onto a plane. Four different encoding schemes were compared regarding their usefulness for proteinase classification: Topological autocorrelation of atom type and accessibility over five (a) and ten (b) distance bins, and topological cross-correlation of atom type and accessibility over five (c) and ten (d) distance bins (Figs. 4–6). We will briefly discuss general trends observed by the application of PCA (Fig. 4) and then proceed to a more thorough discussion of NLM (Fig. 5) and SOM (Fig. 6) results. It should be noted that, in every representation, there are four projections of a thermolysin structure (1tlp). In Figures 4 and 5, they are enclosed by an ellipse. The protein coordinates used for their generation are identical, but assume different spatial orientations relative to the cubic grid that defines the cavity (cf. Fig. 2). This leads to slightly different boundary criteria for the active site cavity, and the corresponding surfaces and correlation vectors also differ to some extent. The area covered by the four points in the different projections may therefore serve to assess the intrinsic error of the method (''internal standard'').

From Figure 4 it appears that a crude clustering of active site types according to functional classes can already be visualized by the two PCA vectors with the largest eigenvalues. In all four cases, they cover about 60% of the total variance in the data set. The serine, aspartic, and metalloproteinases form distinct groups with some interdigitation. A cluster of five dimeric aspartic proteinases (1fiv, 1fmb, 1hii, 1hxb, 1siv) is always separated from their monomeric counterparts. This reflects the differences between the tube-shaped active sites in the homodimeric enzymes and the open active site structures in all other proteinases.

With respect to the separation into proteinase clusters, the subjective rank order of the four plots in Figure 4 is (a) > (b) and (c) > (d). We conclude that the 25-dimensional autocorrelation vectors in Figure 4a already entail essential information about characteristic active site features, which permits clustering of the three classes of proteinases.

Four NLM maps were calculated (Fig. 5). This type of projection yields a significantly clearer picture of the relations between active sites than does PCA. In general, their relative positions in the plots match the expected distribution; that is, the three classes of proteinases form separate clusters. Separation into the three functional classes is better for high-dimensional correlation vectors (Fig. 5b, d). Here, cross-correlation performs slightlybetter than autocorrelation.

In contrary to NLM representations, SOM plots (Fig. 6) do not conserve distances between vectors in the training set. Nevertheless, neighborhood relationships are preserved and data clusters can be

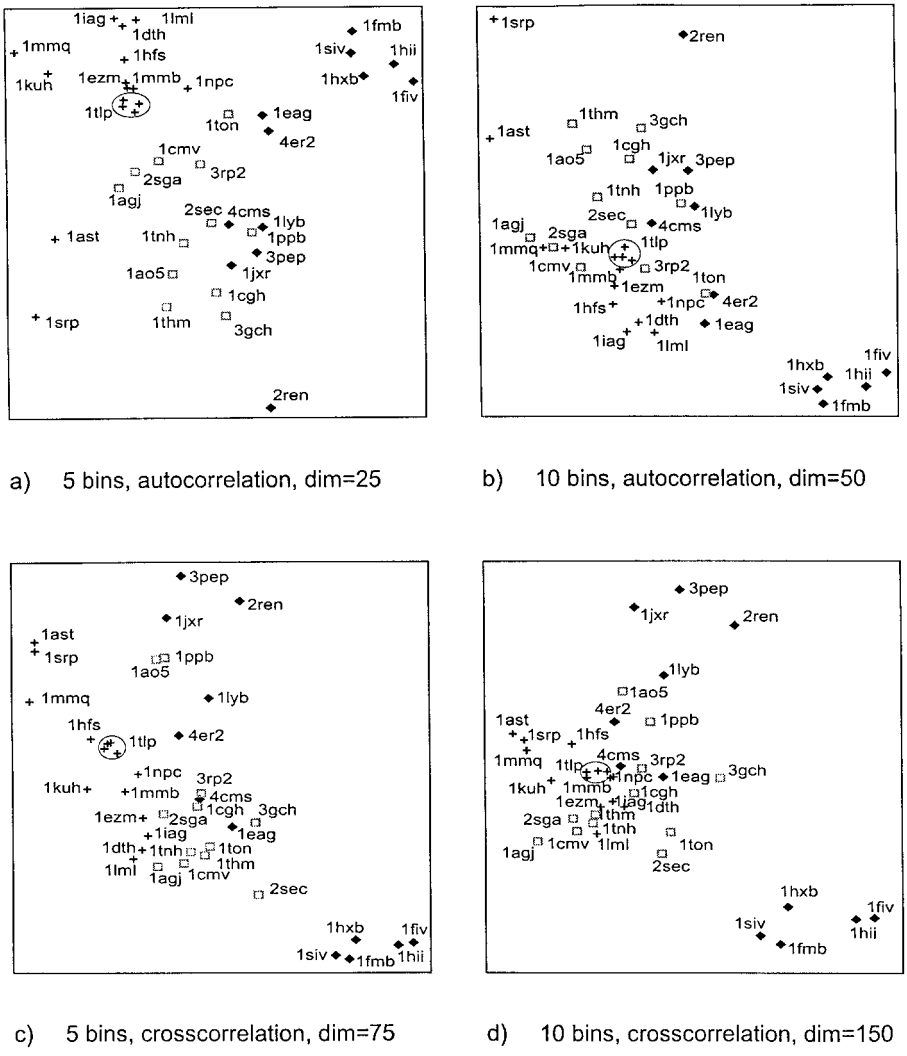

FIGURE 4. Principal component analysis (PCA) of four sets of correlation vectors generated from the data set (Table I) of active site surfaces. The principal components with the largest eigenvalues are plotted against each other. (a) Eigenvalues 8.5, 6.6, covering 60.5% of the total variance; (b) 16.9, 12.9, covering 59.7% of the total variance; (c) 31.4, 15.8, covering 62.9% of the total variance; (d) 61.3, 31.2, covering 61.6% of the total variance. For generation of the correlation vectors, a 10-Å cutoff, divided into the given number of distance bins, was used. "*Dim*" gives the number of components of the resulting vector.

formed by adjacent neurons. For the given number of 36 active sites, the identical number of neurons would have been sufficient to assign one neuron to each vector. To facilitate clustering and convergence of the SOM training process, as well as visual inspection, we used a toroidal map of $10 \times 10 = 100$ neurons. As a result, the group of dimeric aspartic proteinases is entirely surrounded by unoccupied neurons (open squares) in all four SOMs.

Zinc metalloproteinases can be formally divided into a number of subclasses according to their fold. Among them are thermolysin-like proteinases and metzincins. There are three representatives with a thermolysin fold in our selection, 1ezm, 1npc, and 1tlp. As expected, they are grouped closely together on all maps. The two members of the snake venom family—a subclass of the metzincins—adamalysin (1iag) and atrolysin (1dth), are close neighbors in all plots. Other metzincins are grouped together despite rather diverse active site topologies, as can be seen in both NLMs and SOMs [clusters around neuron (3,6) in Fig. 6b, around (7,9) in Fig. 6c, and around (3,2) in Fig. 6d]. Two enzymes of this class (1ast and 1srp) show a

tendency to separate from the bulk of the metalloproteinase group in the NLMs (Fig. 5). Upon scrutinizing their active sites, it is noteworthy that, in contrast to most other metzincins, both have a rather extended C-terminal domain, which forms one side of a narrow and deep active site cleft. Leishmanolysin (1lml) has a very large C-terminal domain as well, but its active site is wider and more shallow compared with the aforementioned structures (1ast and 1srp).

Serine proteinases appear in compact clusters in all four NLMs (Fig. 5). Yet, there are three structures that behave as outliers in all cases: tonin (1ton), thrombin (1ppb), and kallikrein (1ao5). Tonin has a rather distorted active site, due to complexation of zinc by three histidine side chains. His57 (numbering as in chymotrypsin) of the catalytic triad rotates almost 180° around χ^1 and complexes this zinc ion together with His97 and His99 (numbering as in 1ton). As an effect of this complexation, amino acids 214 to 220 are in a very unusual conformation, which hinders access to the S1 pocket for substrates or inhibitors. However, tonin behaves as a true outlier only in the case of autocorrelation. To test whether this behavior was due to an overemphasis of the presence of zinc in the active sites of tonin and the metalloproteinases, the ten-bin autocorrelation vectors are recalculated with the zinc surface points arbitrarily marked as ''acceptor'' instead of ''metal.'' In the resulting NLM and SOM, tonin is again located close to the metalloproteinases (Fig. 7). Obviously,

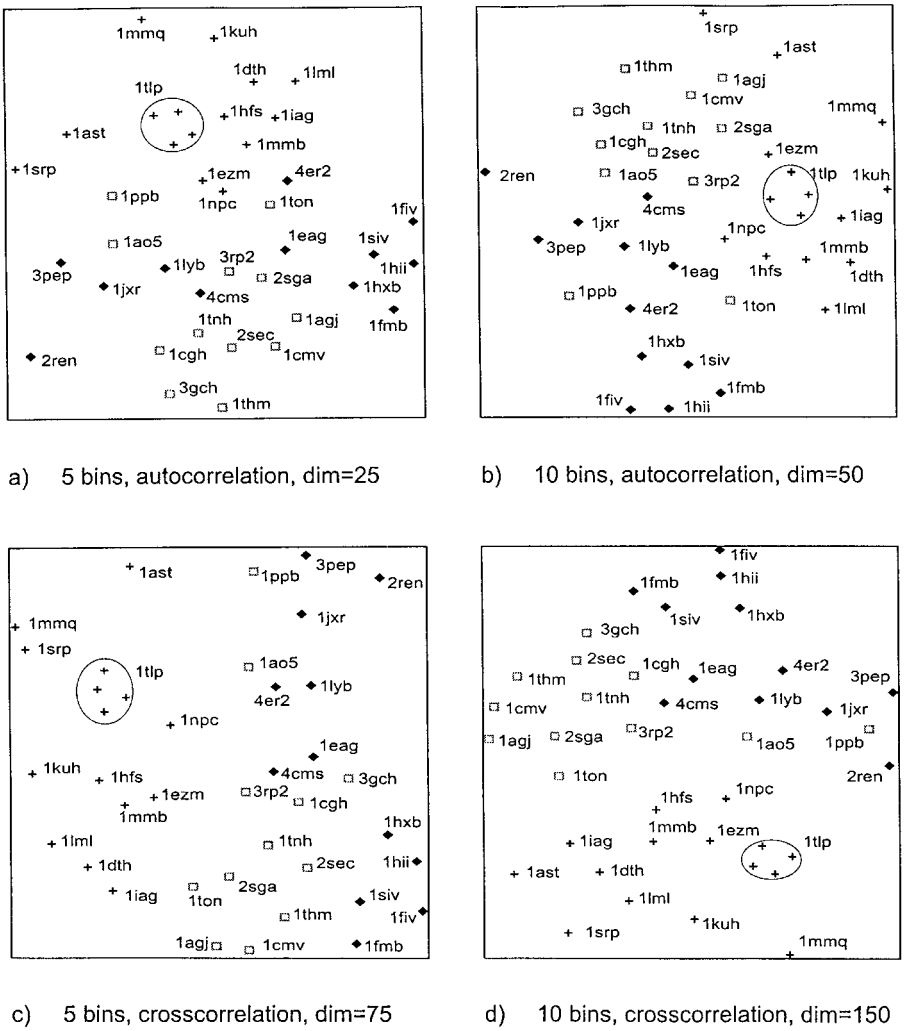

FIGURE 5. Nonlinear maps (NLM) of four sets of correlation vectors generated from the data set (Table I) of active site surfaces. For generation of the correlation vectors, a 10-Å cutoff, divided into the given number of distance bins, was used. ''*dim*'' is the number of components of the resulting vector. Mapping errors are: (a) 0.036; (b) 0.023; (c) 0.016; (d) 0.012.

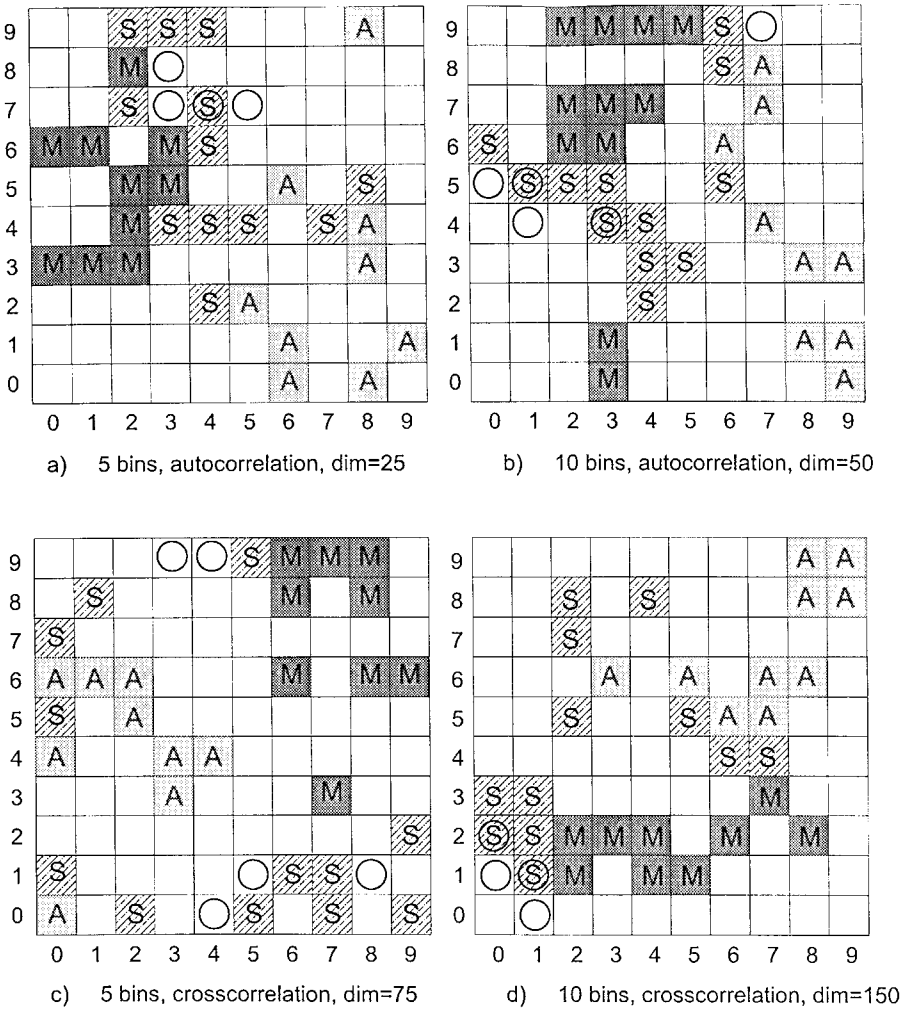

FIGURE 6. Self-organizing maps (SOM) of four sets of correlation vectors generated from the data set (Table I) of active site surfaces. Each neuron is marked by proteinase class. A: aspartic proteinase; S: serine proteinase; M: metalloproteinase. Circles denote occupancy by cysteine proteinases, which are not part of the training set. Note that each map contains 10 × 10 neurons forming a torus. The individual PDB codes are: (0,3): 1npc; (0,6): 1tlp, 1tlp, 1tlp; (1,3): 1dth, 1iag; (1,6): 1tlp; (2,3): 1lml; (2,4): 1hfs, 1mmb; (2,5): 1mmq; (2,7): 1cmv; (2,8): 1ezm; (2,9): 3rp2; (3,4): 1agj; (3,5): 1ast, 1srp; (3,6): 1kuh; (3,7): 1act, 1gec, 1mem, 1pip, 1ppo, 1yal; (3,8): 1meg; (3,9): 2sec; (4,2): 1ao5; (4,4): 1tnh; (4,6): 2sga; (4,7): 1thm, 1ice; (4,9): 3gch; (5,2): 1jxr; (5,4): 1cgh; (5,7): 1the; (6,0): 1lyb; (6,1): 2ren, 3pep; (6,5): 4cms; (7,4): 1ton; (8,0): 1fiv, 1fmb, 1hii; (8,3): 1eag; (8,4): 4er2; (8,5): 1ppb; (8,9): 1siv; (9,1): 1hxb. (b) (0,5): 1meg, 1mem, 1pip, 1ppo, 1yal; (0,6): 1cmv; (1,4): 1gec; (1,5): 2sga, 1act; (2,5): 1agj; (2,6): 1ast, 1srp; (2,7): 1mmq; (2,9): 1dth, 1iag; (3,0): 1tlp; (3,1): 1ezm; (3,4): 1thm, 1ice; (3,5): 1tnh; (3,6): 1kuh; (3,7): 1hfs, 1mmb; (3,9): 1tlp; (4,2): 3rp2; (4,3): 2sec; (4,4): 1cgh; (4,7): 1lml; (4,9): 1tlp; (5,3): 3gch; (5,9): 1npc; (6,5): 1ao5; (6,6): 4cms; (6,8): 1ppb; (6,9): 1ton; (7,4): 1jxr; (7,7): 4er2; (7,8): 1eag; (8,1): 1fiv, 1fmb, 1hii; (8,3): 2ren, 3pep; (9,0): 1hxb; (9,1): 1siv; (9,3): 1lyb; (9,6): 1the. (c) (0,0): 4cms; (0,1): 3gch; (0,4): 1eag; (0,5): 1ton; (0,6): 4er2; (0,7): 1ppb; (1,6): 1jxr; (1,8): 1ao5; (2,0): 3rp2; (2,5): 2ren, 3pep; (2,6): 1lyb; (3,3): 1fmb, 1hii, 1siv; (3,4): 1fiv; (3,9): 1the; (4,0): 1gec, 1mem, 1yal; (4,4): 1hxb; (4,9): 1act, 1pip, 1ppo; (5,0): 2sga; (5,1): 1meg; (5,9): 1agj; (6,1): 1cmv; (6,6): 1dth, 1iag; (6,8): 1kuh; (6,9): 1ast, 1srp; (7,0): 1tnh; (7,1): 1thm; (7,3): 1ezm; (7,9): 1mmq; (8,1): 1ice; (8,6): 1tlp, 1tlp, 1tlp, 1tlp; (8,8): 1lml; (8,9): 1hfs, 1mmb; (9,0): 1cgh; (9,2): 2sec; (9,6): 1npc. (d) (0,1): 1gec, 1ice, 1meg, 1yal; (0,2): 1cmv, 1ppo; (0,3): 1thm; (1,0): 1the; (1,1): 2sga, 1act, 1mem, 1pip; (1,2): 1agj; (1,3): 1tnh; (2,1): 1kuh; (2,2): 1ast, 1srp; (2,5): 1cgh; (2,7): 3gch; (2,8): 2sec; (3,2): 1mmq; (3,6): 4cms; (4,1): 1lml; (4,2): 1hfs, 1mmb; (4,8): 3rp2; (5,1): 1dth, 1iag; (5,5): 1ton; (5,6): 1eag; (6,2): 1tlp, 1tlp, 1tlp, 1tlp; (6,4): 1ppb; (6,5): 4er2; (7,3): 1npc; (7,4): 1ao5; (7,5): 1jxr; (7,6): 2ren, 3pep; (8,2): 1ezm; (8,6): 1lyb; (8,8): 1fiv; (8,9): 1fmb, 1hii; (9,8): 1siv; (9,9): 1hxb.

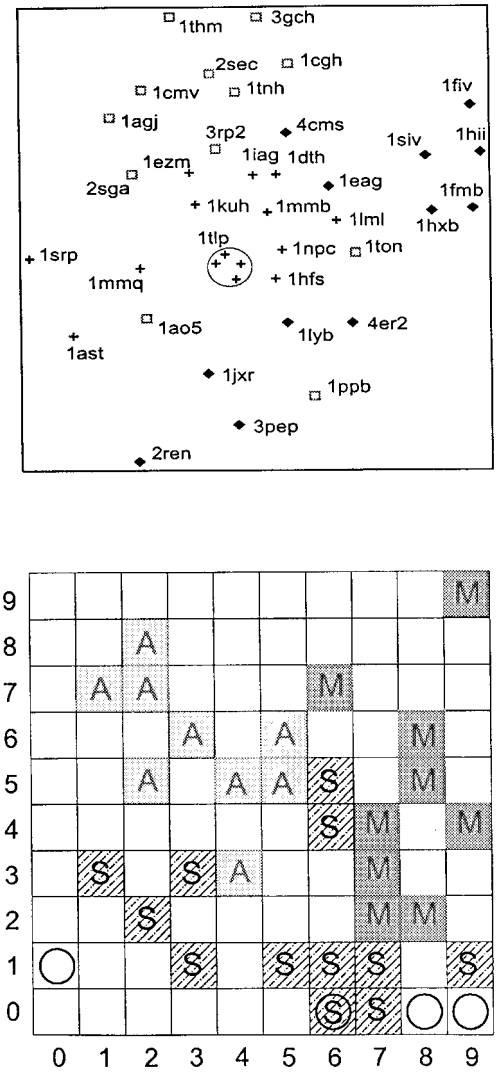

FIGURE 7. NLM and SOM projections of autocorrelation vectors, where metal atoms were arbitrarily treated as H-bond acceptors in all zinc-containing enzymes. In all other respects the vectors were generated as in Figures 3b and 4b, resulting in autocorrelation vectors with 40 components. Note that the SOM contains 10 × 10 neurons forming a torus. Neurons are occupied as follows: (0,1): 1meg, 1the; (1,3): 3rp2; (1,7): 1hxb; (2,2): 2sec; (2,5): 1lyb; (2,7): 1fiv, 1fmb, 1hii; (2,8): 1siv; (3,1): 3gch; (3,3): 1ao5; (3,6): 2ren, 3pep; (4,3): 4cms; (4,5): 1jxr; (5,1): 1cgh; (5,5): 4er2; (5,6): 1eag; (6,1): 1tnh; (6,5): 1ppb; (6,0): 1thm, 1ice; (6,4): 1ton; (6,7): 1npc; (7,0): 2sga; (7,1): 1agj; (7,2): 1ast, 1srp; (7,3): 1mmq; (7,4): 1hfs, 1lml, 1mmb; (8,0): 1act, 1gec, 1pip, 1ppo, 1yal; (8,2): 1kuh; (8,5): 1tlp; (8,6): 1tlp, 1tlp, 1tlp; (9,0): 1mem; (9,1): 1cmv; (9,4): 1dth, 1iag; (9,9): 1ezm.

this is a borderline case in which the two correlation types emphasize the vicinity to the two proteinase classes to a different extent. The plots in Figure 7 also indicate that metalloproteinases form a separate cluster not merely because they contain a zinc ion, but also due to their specific shape and distribution of functional groups in the active site.

At first sight it is rather astonishing to find thrombin (1ppb) being an outlier from the serine proteinase cluster, because there is a number of structures with rather similar overall fold in the chosen set (3gch, 3rp2, 1tnh, 1cgh, and 1ao5). A unique feature in thrombin is a nine-amino-acid insert between residues 60 and 61 which forms a β-hairpin. This rooflike loop covers the S2 pocket and creates a well-defined cavity in thrombin, unlike the cavities in all other serine-proteinases in the set. A similar argument holds for mouse kallikrein (1ao5), which has an 11-amino-acid insert between N95 and G96, compared with chymotrypsin. These additional amino acids form a long loop covering the S2 and S3 pockets. In addition, amino acids 172 to 175 form the distal wall of the S3 pocket in kallikrein. As a consequence, these two pockets are deep and distinctly shaped. This leads to an unusually pronounced cleft in the nonprime region, which is reflected in the location of kallikrein (1ao5) on the maps.

The CMV proteinase (1cmv) differs significantly from all other proteinases in fold and active site shape. It is therefore quite surprising to find it within the appropriate cluster of serine proteases. Its active site is extremely shallow and does not contain the usual Ser–His–Asp catalytic triad (Fig. 1b). Instead, a second His is found in the position of the usual Asp. In all NLMs and SOMs, its next neighbors are *Staphylococcus aureus* toxin A (1agj) and thermitase (1thm), which comprise rather wide and shallow active sites as well. As already found for thrombin, the average accessibility of the active site surface seems to dominate over details of shape.

The aspartic proteinases are split up into two clusters (Figs. 5 and 6). Structures of the homodimeric proteinases (1fiv, 1fmb, 1hii, 1hxb, 1siv) appear as a compact bundle in all NLMs and SOMs. Apart from parts of the flap region, the Cα structures of the dimeric proteinases superimpose very nicely in the active site regions. Even the absence of some residues at the tip of the flaps in two structures (1fmb, 1fiv) does not seem to have a large effect on the active site properties, and thus on clustering behavior of this type of proteinase.

Monomeric members of the aspartic proteinase family are significantly more spread out over the plots. Members of this family comprise rather diverse active sites compared with their dimeric counterparts. The long β-hairpin-like loop formed by amino acids 70 to 83 (numbering as in pepsin), which is usually referred to as a ''flap,'' shows different conformations in all structures of our set. Other regions with conformational heterogeneity in monomeric structures are between amino acids 109 and 116 in the N-terminal domain, whereas there are variations in length and conformation found in regions encompassing amino acids 238 to 245, 275 to 285, and 290 to 298 (pepsin nomenclature). The four NLMs (Fig. 5) nicely reflect these differences in the active site regions of the structures, which exist despite a common overall fold for these monomeric proteinases. Chymosin (4cms), which lacks six amino acids (290 to 296) and comprises a distorted flap, has a rather accessible half-pipe-shaped active site. It can be seen that it is found in close vicinity to the cluster of serine proteinases in all plots. One should also keep in mind that, to a certain extent, the differences among the monomeric aspartic proteinases reflect individual conformations that are enforced by intermolecular interactions in the crystals.

A unique feature of self-organizing maps is that they can be employed to predict the positions of vectors that have not been part of the training set. We have probed the predictive power of the SOMs in Figure 6 by calculating the positions of members of another class of proteinases. A group of nine cysteine proteinases[24] (Table I) was selected for this purpose. In Figure 6, their predicted positions are marked by open circles. Because most of these positions are not occupied by members of the training set, one can conclude that they have distinct active site properties, and our method is able to recognize cysteine proteinases as yet a further class of proteinases. Apart from interleukin-converting enzyme (1ice) they form a rather compact cluster on the maps in close vicinity to serine proteinases. This is a reasonable result, because both serine and cysteine proteinases catalyze the cleavage of amide bonds involving a similar catalytic apparatus (Fig. 1b, c).

Summary and Conclusion

We have developed a novel method for mapping active site cavities in proteins by projecting surface-derived correlation vectors. We are not aware of any published studies of this kind.[25] The method comprises a simple and generally applicable way of defining cavities at the surface of proteins as well as assignment of shape- (accessibility) and functionality-related (generalized atom type) properties to the active site surface. Based on these surfaces, active site descriptors were generated, which suitably code for essential active site features. Two projection methods, nonlinear mapping and self-organizing mapping, proved to be particularly useful for visual data analysis. Applied to a diverse set of aspartic, serine, cysteine, and metalloproteinases, we were able to differentiate between the individual members of three classes of enzymes and to display intra- and interclass relationships. Furthermore, from this analysis, it becomes apparent that the accessibility measure employed for shape description and its use in the correlation vectors overemphasizes size and average accessibility of active sites. A more localized shape descriptor would therefore be desirable. This question as well as an assessment of the general applicability of our method to diverse tasks will be the topic of future work. Future applications will include clustering of active site pockets together with other cavities identified at a protein surface. When this is done for large collections of protein structures, such a procedure could pave the way for a functional classification that is independent from fold type and sequence homology. It could then be employed to localize active sites in proteins in which 3D structure is not yet fully understood. A further challenging application is the use of our correlation vectors as pseudoreceptors guiding the design of enzyme inhibitors.

Acknowledgments

The authors thank our colleagues at Roche for many valuable discussions, especially Hans-Joachim Böhm, Klaus Müller, Chiara Taroni, and Fritz Winkler.

References

1. Levitt, M.; Chothia, C. Nature 1976, 261, 552; Richardson, J. S. Adv Pro Chem 1981, 34, 167; Richardson, J. S.; Richardson, D. C.; Tweedy, N. B.; Gernert, K. M.; Quinn, T. P.; Hecht, M. H.; Erickson, B. W.; Yan, R.; McClain, R. D.; Donlan, M. E.; Surles, M. C. Biophys J 1992, 63, 1186; Thornton, J. M. Curr Opin Struct Biol 1992, 2, 888; Holm, L.; Sander, C. Science 1996, 273, 595; Fischer, D.; Wolfson, H.;

Lin, S. L.; Nussinov, R. Prot Sci 1994, 3, 769; Wallace, A. C.; Borkakoti, N.; Thornton, M. Prot Sci 1997, 6, 2308.

2. Gubernator, K.; Böhm, H.-J. eds.; Structure-Based Ligand Design; Wiley-VCH: Weinheim, 1998. Structure based design methods rely on surface properties of cavities; for instance: Bohacek, R. S.; McMartin, C. J Med Chem 1992, 35, 1671; Blaney, J. M.; Dixon, J. S. Persp Drug Discov 1993, 1, 301.
3. Laskowki, R. A. J Mol Graph 1995, 13, 323; Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B.; Thornton, J. M. Prot Sci 1996, 5, 2438 (and references cited therein).
4. Rosen, M.; Lin, S. L.; Wolfson, H.; Nussinov, R. Prot Eng 1998, 11, 263 (and references cited therein).
5. Dean, P. M., ed. Molecular Similarity in Drug Design; Blackie: Glasgow, 1995; Balaban, A. T., ed. From chemical Topology to Three-Dimensional Geometry; Plenum: New York, 1997.
6. Bernstein, F. C.; Koetzle, T. E.; Williams, G. J. B.; Meyer, Jr., E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112, 535.
7. The evenly distributed sphere points were obtained from Jon Leech, University of North Carolina, Chapel Hill, via ftp (ftp.cs.unc.edu/pub/users/leech/points.tat.gz). For a discussion on the topic, see Saff, E. B.; Kuijlaars, A. B. J. Math Intell 1997, 29, 5.
8. Connolly, M. L. Science 1983, 221, 709; Connolly, M. L. J Appl Cryst 1983, 16, 548.
9. Sadowski, J.; Wagener, M.; Gasteiger, J. Angew Chem Int Ed Engl 1995, 34, 2674.
10. Wagener, M.; Sadowski, J.; Gasteiger, J. J Am Chem Soc 1995, 117, 7769.
11. Bienfait, B.; Gasteiger, J. J Mol Graph Model 1997, 15, 203; Domine, D.; Devillers, J.; Chastrette, M.; Karcher, W. J Chemometrics 1993, 7, 227.
12. Schneider, G.; Wrede, P. Prog Biophys Mol Biol (in press).
13. Jollife, I. T. Principal Component Analysis; Springer: New York, 1996.
14. TSAR 3.1, Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Oxford, UK.
15. Sammon, Jr., J. W. IEEE Trans Comput 1969, C-18, 401.
16. Rechenberg, Evolutionsstrategie-Optimierung technischer Systeme nach Prinzipien der biologischen Evolution; Fromann-Holzboog: Stuttgart, 1973.
17. Schneider, G.; Schuchhardt, J.; Wrede, P. Biol Cybernet 1996, 74, 203.
18. Kohonen, T. Biol Cybernet 1982, 43, 59; Kohonen, T. Self-Organization and Associative Memory; Springer: Heidelberg, 1989.
19. Hertz, J.; Krogh, A.; Palmer, R. G. Introduction to the Theory of Neural Computation; Addison-Wesley: Redwood City, CA, 1991.
20. Zupan, J.; Gasteiger, J. Neural Networks for Chemists; VCH: Weinheim, 1993.
21. Davies, D. R. Annu Rev Biophys Biophys Chem 1990, 19, 189; Guruprasad, K.; Dhanaraj, V.; Groves, M.; Blundell, T. L. Persp Drug Discovery Des 1994, 2, 329.
22. Fersht, A. Enzyme Structure and Mechanism, 2nd Ed.; W. H. Freeman: New York, 1985; p 405.
23. Stocker, W.; Bode, W. Curr Opin Struct Biol 1995, 5, 383; Stocker, W.; Grams, F.; Baumann, U.; Reinemer, P.; Gomis-Ruth, F. X.; McKay, D. B.; Bode, W. Prot Sci 1995, 5, 823.
24. Otto, H.-H.; Schirmeister, T. Chem Rev 1997, 97, 133.
25. During the final stages of this project, we became aware that similar work is in progress in the group of Prof. G. Klebe, Marburg (personal communication).